

基于 LSTM 的商品评论情感分析^①

於 雯, 周武能

(东华大学 信息科学与技术学院, 上海 201620)

摘要: 随着电子商务的发展, 产生了大量的商品评论文本. 针对商品评论的短文本特征, 基于情感词典的情感分类方法需要大量依赖于情感数据库资源, 而机器学习的方法又需要进行复杂的人工设计特征和提取特征过程. 本文提出采用长短期记忆网络 (Long Short-Term Memory) 文本分类算法进行情感倾向分析, 首先利用 Word2vec 和分词技术将评论短文本处理为计算机可理解的词向量传入 LSTM 网络并加入 Dropout 算法以防止过拟合得出最终的分类模型. 实验表明: 在基于深度学习的商品评论情感倾向分析中, 利用 LSTM 网络的短时记忆独特特征对商品评论的情感分类取得了很好的效果, 准确率达到 99% 以上.

关键词: 情感分析; 商品评论; 长短期记忆网络; 自然语言处理; 深度学习

引用格式: 於雯, 周武能. 基于 LSTM 的商品评论情感分析. 计算机系统应用, 2018, 27(8): 159-163. <http://www.c-s-a.org.cn/1003-3254/6483.html>

Sentiment Analysis of Commodity Reviews Based on LSTM

YU Wen, ZHOU Wu-Neng

(School of Information Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: With the development of e-commerce, a large number of reviews of goods have been produced. According to short text features of commodity reviews, emotion classification based on sentiment dictionary needs a lot of emotion database resources, and machine learning method needs complex artificial design features and feature extraction process. This study adopts the short and long term memory network (Long Short Term Memory, LSTM) text classification algorithm sentiment analysis. Firstly, the text word vector is introduced into LSTM network by using Word2Vec and word segmentation technology, and finally the classification model is obtained by Dropout algorithm. Experiments show that: in deep learning based sentiment analysis of commodity reviews, the unique characteristics of short-term memory network have sound results on commodity reviews sentiment classification, the accuracy of classification is more than 99%.

Key words: sentiment analysis; product review; Long Short-Term Memory (LSTM); natural language processing; deep learning

引言

互联网的飞速发展改变了人民的生活方式给人们的生活带来了很大的便利, 也极大地推动了电子商务的发展, 网络购物也逐渐进入人们的生活中. 但是网络购物在给人们带来便捷的同时也具有很多不可避免的弊端, 无法真实的感受到商品的属性, 就为我们购买商

品带来了很大的风险, 所以我们经常都会查看别人留下的评论信息在我们做出购买商品决策之前来提供参考. 各大网络购物平台也会要求人们在购物结束之后留下商品的评论信息, 为商家提供生产指导也为后续的消费者提供参考. 对商品评论信息的情感倾向分析不仅能够为消费者购买商品是提供指导, 也可以使商

① 基金项目: 国家自然科学基金 (NSFC61573095)

Foundation item: National Natural Science Foundation of China (NSFC61573095)

收稿时间: 2017-12-18; 修改时间: 2018-01-04; 采用时间: 2018-01-16; csa 在线出版时间: 2018-07-28

家客观的认识到自己商品的优缺点,来改进自己的商品.因此对商品评论的文本情感分析不仅具有巨大的商业上的价值也可以对学术研究用到实际领域起到很好的推动作用^[1].

1 相关工作

目前,对于商品评论情感倾向分类使用较多的方法主要有基于情感词典和基于传统机器学习方法.基于情感词典的方法^[2],主要依赖于对大量情感词语数据库的构建,而且现有的情感词典规模有限并且情感词对于所在领域和属性的不同具有很大的区别,因此该方法的应用具有较大的局限性.传统机器学习的方法最主要思想是先要通过一部分预先标注好的文本数据进行预训练,训练出一个情感倾向分类器,然后通过不断的调节参数修改模型以达到较好分类性能,再通过这个训练好的情感分类器对所有的文本进行情感倾向极性的分类.如:针对不同的特征选择方法,分别应用朴素贝叶斯、支持向量机(SVM)、最大熵等不同的机器学习的分类方法.对评论进行分类,在几种分类方法中SVM分类方法取得的效果最好,其分类准确率最高可以达到约88%^[3,4].

近年来,深度学习的算法在语音和图像领域取得较大突破,将深度学习的方法应用与自然语言处理领域也获得了较大的关注.如Bengio等利用深度学习的思想构建的神经概率语言模型^[5],将各种深层神经网络在大规模英文语料上进行语言模型训练,得到了较好的语义表征,完成了句法分析和命名实体识别等许多的自然语言处理任务,使得自然语言处理任务也能很好的应对大数据时代的挑战^[6,7].本文采用深度学习的循环神经网络中的一种特殊的长短期记忆网络LSTM来对商品评论进行情感倾向分析.

2 分类模型

2.1 词向量表示

在文本处理中,首要的问题就是将人类的语言文本转化为计算机可以理解的计算机语言,就是如何把一个句子用矢量的形式有效地表达出来.如果这一步能够很好的完成,文本分类的复杂程度将会被大大降低.首先可以将每个词语都用一个唯一的编号来表示,然后把每个词的编号集合起来就变成一个句子的表示,并且对于词语的语义相似性问题,可以通过对语义相似的词语赋予相近的编号就可以很好的在保证语义的基础上的表示出一个句子,实际上这种方法只在假定

语义是单一的情况下成立,也就是说,在这种情况下考虑的语义是只有一维的情况.但是事实却并非如此简单,语义应该是更加复杂的多维的情况,而且基于中文的语义就会更加庞杂.因此如果只通过一个唯一的编号,就想将这些词语放到合适的位置具有很大的难度.针对上述问题本文利用Google开源的著名工具Word2Vec^[8]用高维的向量(词向量, Word Embedding)来表示词语,并把具有相近意思的词语放在同样相近的位置并且对固定词向量的维度,就可以通过实数向量来训练模型,以获得词语的词向量表示.

2.2 LSTM网络

循环神经网络(Recurrent Neural Network, RNN)^[9],顾名思义是具有循环结构的神经网络,其网络结构如图1所示.RNN的链式结构使得其能够很好的解决序列的标注问题.但是由于梯度爆炸和梯度消失等问题经常存在于在RNN的训练中,使得RNN很难保持较长时间的记忆.

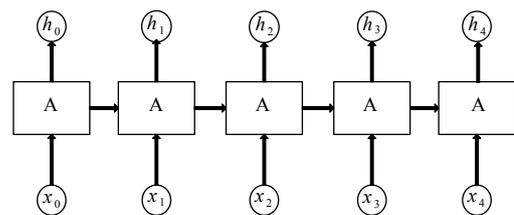


图1 RNN网络结构

长短期记忆(Long Short-Term Memory, LSTM)网络是有RNN扩展而来,在设计之初就是用来解决RNN长期依赖的问题.LSTM也是一种特殊的循环神经网络因此也具有链状结构,但是相比循环神经网络的重复模块有着不同的结构.它有四层神经网络层,各个网络层之间以特殊的方式相互作用,并非单个简单的神经网络层.网络模块示意图如图2所示.

在图2中每个传输单元的状态是决定LSTM网络的核心,就是穿过图中的每条水平线.一个单元状态就相当于于一个传送带,它贯穿于整个结构,在这个过程中仅通过一些线性的作用保证了信息传输的不变性.LSTM还具有一个很好的性能就是可以增加和去除向单元状态中传输的信息,通过几个结构来管理信息的传输并将其称为门限.门限就是有选择地让信息通过.它们由一个sigmoid神经网络层和逐点乘法运算组成.sigmoid层输出[0, 1]的数字,描述了每个成分应该通过门限的程度.0表示不让任何成分通过,而1表示所有成分都可以通过^[10].

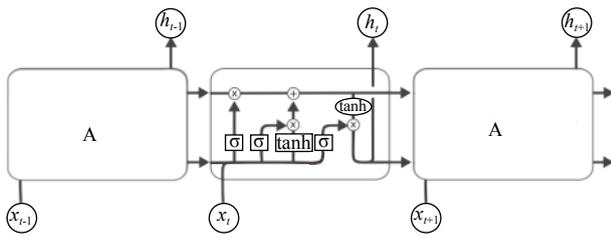


图2 LSTM网络结构

LSTM有三种这样的门限,分别为:输入门、遗忘门和输出门通过这三个门限来保护和控制单元状态.以下公式中 i_t , f_t , o_t 和 c_t 分别表示 t 时刻对应的三种门限结构和单元状态.

(1) 遗忘状态信息, LSTM中第一步是决定从上一单元传来的信息哪些需要从单元状态中抛弃.这是由一个sigmoid输出层决定的该层被成为“遗忘门”.然后为单元状态中的每个数字输出一个[0, 1]的值.就是以当前层的输入 x_t 和上一层的输出 h_{t-1} 作为输入,然后在 $t-1$ 时刻的细胞状态输出 f_t ;

$$f_t = \sigma(W_f \cdot (h_{t-1}, x_t) + b_f) \quad (1)$$

(2) 更新状态信息, 决定在单元状态中哪些新信息需要被存储下来.这一层可以分为两部分.首先,由一个sigmoid层决定哪些值需要更新该层被称为“输入门”.接下来, tanh层创建一个向量,里面包含新的待添加信息,这些信息是可以添加到新的状态单元的信息.

下一步我们将会结合这两部分的信息对单元状态进行更新.首先更新旧单元状态,然后为新的单元状态输入新的信息,即在细胞状态中更新信息,主要传输过程为:
① 输入门的sigmoid层的结果 i_t 作为将更新的信息; ② 由tanh层新创建的向量 \tilde{c}_t ,将添加到在细胞状态中.将旧的细胞状态 c_{t-1} 乘以 f_t ,用以遗忘信息,与新的候选信息 $i_t \cdot \tilde{c}_t$ 的和,一起合成细胞状态的更新.

$$i_t = \sigma(W_i \cdot (h_{t-1}, x_t) + b_i) \quad (2)$$

$$\tilde{c}_t = \tanh(W_c \cdot (h_{t-1}, x_t) + b_c) \quad (3)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \quad (4)$$

(3) 输出信息,用sigmoid层来决定要输出细胞状态的相关信息,该层被称为“输出门”然后用tanh处理细胞状态,两部分信息的乘积就是要输出的信息.

$$o_t = \sigma(W_o \cdot (h_{t-1}, x_t) + b_o) \quad (5)$$

$$h_t = o_t \cdot \tanh(c_t) \quad (6)$$

其中, sigmoid函数模型为:

$$\sigma(x) = (1 + e^{-x})^{-1} \quad (7)$$

Tanh函数模型为:

$$h(t) = \sigma(W_1 x(t) + b_1) \quad (8)$$

2.3 情感分类模型设计

基于LSTM我们可以搭建出商品评论文本的情感分类模型,其结构图如图3所示.

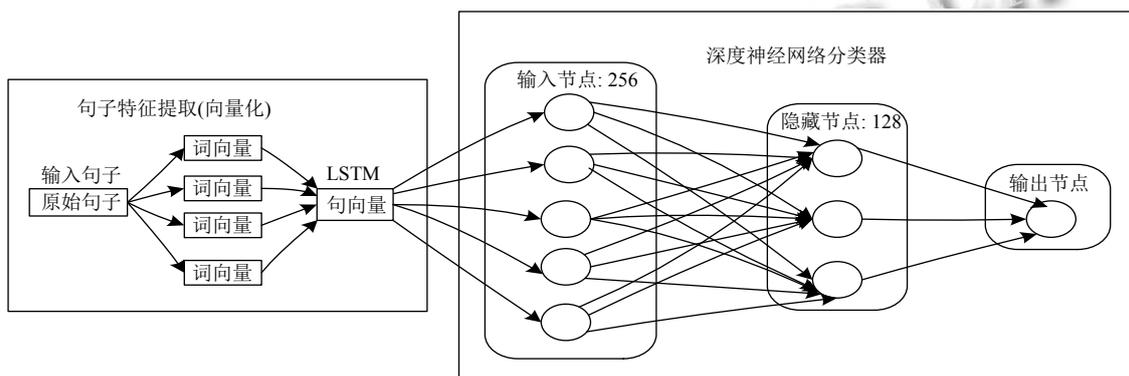


图3 分类模型结构图

由于互联网的商品评论大多属于短文本,且每个用户的语言习惯具有很大差异,在进行文本分类之前需要对文本进行预处理,需要去除大量的无关符号,同时对词频进行统计,统计得出在文本中出现频率最高的多为与情感表达无关的副词,因此利用分词工具进行词性标注后去除无关的停用词.将经过预处理后的

文本作为输入,进入分类模型之后通过 Word2vec 把文本变为分布式存储的词向量,经过 LSTM 网络,为防止模型产生过拟合,在 LSTM 网络中加入 Dropout 随机优化因子,使得模型在复杂的评论语句中也能很好的拟合,增加了模型的适用性

3 实验结果与分析

本文的实验环境为 Windows 7 操作系统, 采用 Python 编程语言, 利用 Python 的 Keras 库来实现 LSTM 网络的搭建, 其中 Keras 库底层通过调用 Tensorflow 框架来实现本文实验环境的搭建, 对比实验中的机器学习方法, 也采用 Python 的第三方机器学习库 Scikit-learn 来实现传统机器学习的分类方法。

3.1 实验数据

本文实验的英文数据采用斯坦福大学公开的评论数据^[12], 数据已做好情感分类标签正面情感标为 1 负面情感标为 0。中文数据利用网络爬虫工具从网站上爬取两万多条评论信息(涉及六个领域)在去除无关信息后进行人工标注。在标注过程中由三个人分别于不同时间段进行标注正面情感标为 1 负面情感标为 0, 并将标注结果取三个人中的最大值(即如果对某条信息的标注结果不同则取标注结果人数最多的作为最后结果)以保证标注结果的准确性。对中文实验数据采用 jieba 分词工具做中文分词处理并去除停用词。最后采用随机切分的方法随机将数据切分为测试集和训练集, 测试集和训练集的占比为 2:8, 训练集用于训练 LSTM 模型, 测试集用于测试分类模型的分类效果。

3.2 评价指标

实验采用的评估分类模型的主要指标为模型分类的准确率、召回率及 F1 值^[11], 具体计算公式如下:

$$precision = \frac{a}{a+b} \times 100\% \quad (9)$$

$$recall = \frac{a}{a+c} \times 100\% \quad (10)$$

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \quad (11)$$

其中, a 表示情感分类模型判断文本类别正确的数目; b 表示情感分类模型判断错误将本不属于此类的文本分到本类别的数目; c 表示分类器错误判断文本本该属于次类别但却错分为其他类别的数目。

3.3 模型参数

在本文实验模型中主要参数有模型对预测结果进行分类的阈值大小, 还有采用 Dropout 算法的比率, 以及每轮训练数据的大小 (batch_size) 和训练轮数, 模型参数设置如表 1 所示。

由于在对最后的预测结果输出的是一个 [0, 1] 之间的数值所以分类的阈值对于分类的准确率具有很大的影响, 通常会选择中间值 0.5 作为分类的阈值, 就

是在分类结果大于 0.5 时判断为正面, 预测结果小于 0.5 时判断为负面情感, 但是将 0.5 作为阈值有时并不是最好的选择, 如图 4 所示, 在研究分类阈值对于真正率和真负率的影响时发现, 曲线在 0.39 左右发生了陡变, 因此我们选择 0.39 作为模型的分类阈值, 大于 0.39 的才被分为正面情感, 小于 0.39 的就被认为是负面情感。

表 1 基于 LSTM 的模型参数设置

参数	参数值
阈值	0.39
Dropout	0.5
batch	64
迭代次数	10

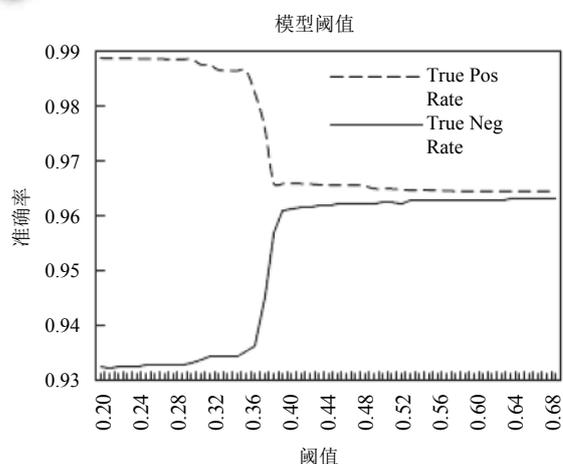


图 4 真正率和真负率随阈值变化曲线

3.4 对比实验分析

本文的方法采用了深度学习的 LSTM 网络利用高维词向量的 Word2vec 方法表示文本词向量以及使用 Dropout 的算法在收集的评论文本中进行实验, 并和其他方法进行对比实验。

将本文采用的 Dropout 算法进行对比实验。因为本文属于二分类问题所以采用对数损失 (binary loss) 来衡量损失, 本文的实验模型中引用了 Dropout 的算法, 从图 5 中可以看到相比于没有使用 Dropout 的算法的模型, 使用 Dropout 的算法之后损失函数开始下降较快但是后期下降比较平稳说明了函数的拟合情况较好, 证明 Dropout 算法的引用提高了模型的泛化能力有效的防止过拟合, 并且可以看到随着训练轮数的增加函数损失不断平稳减小。

最后实验将本文采用的基于 LSTM 的方法与传统的机器学习算法, 朴素贝叶斯和支持向量机的方法和

基于 Word2vec 情感词典 (本文采用的情感词典为知网 HowNet 情感词典) 的方法进行建模对比,同时在中英文数据集上进行验证实验,实验结果如表 2 所示在利用网络爬虫和自己进行人工标注的数据上采用本文 LSTM + Word2vec+Dropout 算法分类模型的各项评价指标都要高于其他几种方法.采用机器学习方法由于需要进行特征的选择和降维操作仍需要较大的工作量,基于情感词典的方法由于对传统的情感词典具有较大的依赖性,并且由于语言系统的复杂性想要进一步提高精度仍具有很大的难度.由此可以证实本文提出的方法对于商品评论情感倾向性分析具有较好的实用性.

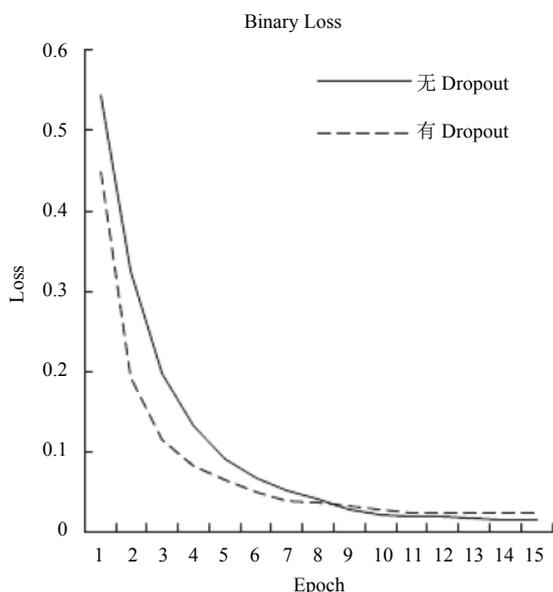


图 5 Dropout 算法对损失函数影响曲线

表 2 不同数据集不同模型对比实验结果

情感分类模型	中文数据集			斯坦福数据集		
	准确率	召回率	F1 值	准确率	召回率	F1 值
Naive Bayesian	0.861	0.834	0.859	0.877	0.858	0.850
Svm	0.884	0.855	0.860	0.891	0.829	0.845
Word2vec+情感词典	0.862	0.833	0.842	0.875	0.846	0.852
LSTM+Word2vec+ Dropout	0.991	0.992	0.998	0.984	0.990	0.987

4 结论与展望

本文的工作基于 LSTM 网络模型,利用 jieba 分词工具对文本语料进行中文分词,并利用 Word2vec 工具来表征高维词向量,训练多领域的文本语料易于移植到多领域的使用克服了传统情感词典方法的领域词典

不足的情况.和传统机器学习的分类模型相比,基于 LSTM 模型的方法,不用人为地设计和提取特征,传统机器学习方法对于文本特征的配置和提取都具有一定的难度,并且通过 Word2vec 对词向量的处理降低了词语的维度同时也缩短了分词模型的训练时间,同时由于深度学习的模型还可以通过 GPU 来训练,可以缩短相应的训练时间,模型的适用性更强.为后续可以很好的使用评论信息来为预测用户喜好以及分析真实商品信息提供了良好的基础支持.

参考文献

- 1 吴燕波, 向大为, 麦永浩. 大数据时代的网络舆情管理与引导研究. 信息安全研究, 2016, 2(4): 356-360.
- 2 李涵昱, 钱力, 周鹏飞. 面向商品评论文本的情感分析与挖掘. 情报科学, 2017, 35(1): 51-55.
- 3 Zou H, Tang XH, Xie B, et al. Sentiment classification using machine learning techniques with syntax features. Proceedings of 2015 International Conference on Computational Science and Computational Intelligence. Las Vegas, NV, USA. 2015. 175-179.
- 4 魏广顺, 吴开超. 基于词向量模型的情感分析. 计算机系统应用, 2017, 26(3): 182-186. [doi: 10.15888/j.cnki.csa.005655]
- 5 Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.
- 6 Bengio Y, Courville A, Vincent P. Representation Learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828. [doi: 10.1109/TPAMI.2013.50]
- 7 Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. arXiv:1301.3781, 2013.
- 8 黄仁, 张卫. 基于 word2vec 的互联网商品评论情感倾向研究. 计算机科学, 2016, 43(6A): 387-389. [doi: 10.11896/j.issn.1002-137X.2016.6A.092]
- 9 Zheng XQ, Chen HY, Xu TY. Deep learning for Chinese word segmentation and POS tagging. Proceedings of 2013 Conference on Empirical Methods in Natural Language, Seattle, WA, USA. 2013. 647-657.
- 10 任智慧, 徐浩煜, 封松林, 等. 基于 LSTM 网络的序列标注中文分词法. 计算机应用研究, 2017, 33(5): 1321-1324, 1341. [doi: 10.3969/j.issn.1001-3695.2017.05.009]
- 11 李胜宇, 高俊波, 许莉莉. 面向酒店评论的情感分析模型. 计算机系统应用, 2017, 26(1): 227-231. [doi: 10.15888/j.cnki.csa.005511]
- 12 梁军, 柴玉梅, 原慧斌, 等. 基于极性转移和 LSTM 递归网络的情感分析. 中文信息学报, 2015, 29(5): 152-159. [doi: 10.3969/j.issn.1003-0077.2015.05.020]